

**PREDICTIVE FRAMEWORK FOR IMBALANCE DATASET**

**MEGAT NORULAZMI BIN MEGAT MOHAMED NOOR**

**DOCTOR OF PHILOSOPHY  
UNIVERSITI UTARA MALAYSIA  
2012**



Awang Had Salleh  
Graduate School  
of Arts And Sciences

Universiti Utara Malaysia

**PERAKUAN KERJA TESIS / DISERTASI**  
(Certification of thesis / dissertation)

Kami, yang bertandatangan, memperakukan bahawa  
(We, the undersigned, certify that)

**MEGAT NORULAZMI MEGAT MOHAMED NOOR**

calon untuk Ijazah  
(candidate for the degree of)

**PhD**

telah mengemukakan tesis / disertasi yang bertajuk:  
(has presented his/her thesis / dissertation of the following title):

**"PREDICTIVE FRAMEWORK FOR IMBALANCE DATASET"**

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.  
(as it appears on the title page and front cover of the thesis / dissertation).

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : **04 Mei 2011**.

*That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on:*  
**May 04, 2011.**

Pengerusi Viva:  
(Chairman for VIVA)

**Prof. Dr. Zulkhairi Md Dahalin**

Tandatangan  
(Signature)

Pemeriksa Luar:  
(External Examiner)

**Assoc. Prof. Dr. Adnan Hassan**

Tandatangan  
(Signature)

Pemeriksa Dalam:  
(Internal Examiner)

**Dr. Yuhanis Yusof**

Tandatangan  
(Signature)

Nama Penyelia/Penyelia-penyelia:  
(Name of Supervisor/Supervisors)

**Assoc. Prof. Dr. Shaidah Jusoh**

Tandatangan  
(Signature)

Tarikh:  
(Date) **May 04, 2011**

## **Permission to Use**

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences  
UUM College of Arts and Sciences  
Universiti Utara Malaysia  
06010 UUM Sintok

## Abstrak

Sasaran penyelidikan ini ialah untuk mengenalpasti dan mencadangkan satu kerangka penyenggaraan jangkaan baharu yang boleh digunapakai dalam menghasilkan satu model ramalan untuk ragam sifat-sifat bahan pemprosesan. Data kadar produksi sebenar yang diperoleh daripada Fuji Electric Malaysia telah digunapakai dalam penyelidikan ini. Penyelidikan ini menggunakan kaedah yang telah diubahsuai daripada konsep prapemprosesan data dan kaedah klasifikasi yang sedia ada. Elemen-elemen kerangka cadangan ialah; membina satu kaedah untuk mengkorelasikan kecacatan bahan pemprosesan, membina satu kaedah untuk mewakili ciri atribut data, menganalisis pelbagai nisbah dan jenis persampelan semula, menganalisis kesan pengurangan dimensi data pada pelbagai saiz data, saiz pembahagian data dan juga skema algoritma terhadap prestasi ramalan. Hasil eksperimen mencadangkan bahawa taburan kebarangkalian kelas untuk model ramalan perlu hampir pada data latihan. Sekitaran kelas data seimbang membolehkan skema algoritma menemukan fungsi  $F$  yang lebih baik daripada ruang  $F_{all}$  yang lebih besar di dalam ruangan ciri berdimensi tinggi, dan *precision* dan *recall* tampak meningkat secara berkadar terus dengan saiz persampelan dan pembahagian data jika nisbah taburan kelas adalah seimbang. Hasil kajian perbandingan yang telah dijalankan juga menunjukkan kaedah yang dicadangkan berprestasi lebih baik. Penyelidikan ini telah dijalankan berdasarkan pada jumlah set data, set data ujian dan pembolehubah yang terhad. Oleh itu, hasil yang diperolehi hanyalah boleh digunapakai pada domain kajian dengan set data yang dikumpul. Penyelidikan ini telah memperkenalkan satu kerangka penyenggaraan jangkaan baharu yang boleh digunapakai oleh industri pembuatan dalam menghasilkan satu model ramalan berdasarkan pada ragam sifat-sifat bahan pemprosesan. Dengan ini, ia membolehkan industri pembuatan menjalankan aktiviti-aktiviti penyelenggaraan jangkaan tidak hanya untuk alatan sahaja tetapi untuk bahan-bahan pemprosesan juga. Sumbangan utama penyelidikan ini ialah garis panduan yang terdiri daripada langkah-langkah kaedah/pendekatan dalam menghasilkan satu model ramalan untuk bahan-bahan pemprosesan.

**Kata kunci:** Penyelenggaraan rangka kerja ramalan, Ketidakseimbangan klasifikasi set data, Pensampelan semula data, Prapemprosesan data, Rangka kerja ramalan.

## Abstract

The purpose of this research is to seek and propose a new predictive maintenance framework which can be used to generate a prediction model for deterioration of process materials. Real yield data which was obtained from Fuji Electric Malaysia has been used in this research. The existing data pre-processing and classification methodologies have been adapted in this research. Properties of the proposed framework include; developing an approach to correlate materials defects, developing an approach to represent data attributes features, analyzing various ratio and types of data re-sampling, analyzing the impact of data dimension reduction for various data size, and partitioning data size and algorithmic schemes against the prediction performance. Experimental results suggested that the class probability distribution function of a prediction model has to be closer to a training dataset; less skewed environment enable learning schemes to discover better function  $F$  in a bigger  $F_{all}$  space within a higher dimensional feature space, data sampling and partition size is appear to proportionally improve the precision and recall if class distribution ratios are balanced. A comparative study was also conducted and showed that the proposed approaches have performed better. This research was conducted based on limited number of datasets, test sets and variables. Thus, the obtained results are applicable only to the study domain with selected datasets. This research has introduced a new predictive maintenance framework which can be used in manufacturing industries to generate a prediction model based on the deterioration of process materials. Consequently, this may allow manufactures to conduct predictive maintenance not only for equipments but also process materials. The major contribution of this research is a step by step guideline which consists of methods/approaches in generating a prediction for process materials.

**Keywords:** Predictive maintenance framework, Imbalanced dataset classification, Data re-sampling, Data pre-processing, Predictive framework

## **Acknowledgement**

Alhamdulillah, praise be to Allah, the Most Gracious, and the Most Merciful. First of all, I would like to thank my supervisor, Dr. Shaidah Jusoh, who saw value in my research and agreed to supervise it. She is very committed in providing support and giving directions for the research. Being a remote PhD student working at Fuji Electric Malaysia, previously and then now with Yayasan Ilmuwan, it is a special situation and I am really grateful that such an arrangement was possible. Furthermore, I am really grateful with her encouragement, effort, and proud on obtaining the Fundamental Research Grant Scheme (FRGS) from MOHE. For all this and more, thank you, Dr. Shaidah Jusoh.

I would also like to thank my previous friends and colleagues: Warikh, Effendy, Monier, Abdullah, Azman, Salahuddin, Jasni, Rasdan, Teo, John, Chew, Natori and others at Fuji Electric Malaysia for being great friends and, who provided an excellent and stimulating working environment and always had time for interesting discussions, arguments, and many interesting ideas that got born this way. Not to forget to thank my current employer, Yayasan Ilmuwan top management (Dr. Khairul and En. Mazilan), who believing in what I am doing by sponsoring few conferences trips. I would also like to thank my Yayasan Ilmuwan friends and colleagues: Zahari, Khairul and Azwan

I am also grateful to the friendly people who volunteered to read through, and give me invaluable comments on this dissertation and its earlier versions, my supervisor Dr. Shaidah Jusoh and Dr. Anitawati Mohd Lokman, Senior Lecturer of UiTM Shah Alam. Without your help this thesis would not have gotten to this stage. Clearly, I am solely responsible for any mistakes that had remained in this thesis. Last but not least, I am deeply indebted to my family for their everlasting support and having by far more faith in me than anybody else, including myself! My special thanks go to Norasikin for being the best wife and a wonderful life companion and for putting up with me and children (Syahmi, Nadiah, Najihah and Farhan) during the hectic time while working on my PhD.

## Table of Contents

Permission to Use .....	ii
Abstrak .....	iii
Abstract .....	iv
Acknowledgement .....	v
Table of Contents.....	vi
List of Tables.....	x
List of Figures.....	xiii
List of Appendices.....	xviii
List of Abbreviations .....	xix
<b>CHAPTER ONE INTRODUCTION .....</b>	<b>1</b>
1.1 Research Background.....	1
1.2 Problem Statement .....	4
1.3 Research Goals .....	5
1.4 Research Questions .....	5
1.5 Objectives .....	6
1.6 Research Approach .....	7
1.7 Definition of Terms.....	12
1.8 Research Scope and Limitations.....	12
1.9 Research Contribution.....	14
1.9.1 Methodological Contributions .....	14
1.9.2 Theoretical Contribution.....	14
1.9.3 Experimental Contribution .....	15
1.10 Publications of the Research.....	15
1.11 Other contributions .....	15
1.12 Thesis Organization .....	15
1.13 Conceptual Framework .....	18
<b>CHAPTER TWO LITERATURE REVIEW.....</b>	<b>20</b>
2.1 Overview .....	20
2.2 Predictive Maintenance .....	22

2.3 Existing Predictive Maintenance Framework.....	29
2.4 Statistical Learning .....	33
2.5 Feature Learning .....	36
2.6 Time-series data transformation .....	39
2.7 Imbalance class distribution dataset.....	41
2.8 Ensemble Learning .....	47
2.9 Summary .....	50
<b>CHAPTER THREE RESEARCH METHODOLOGY .....</b>	<b>52</b>
3.1 Overview .....	52
3.2 Classification methodology .....	53
3.3 Evaluation Measurement.....	59
3.3.1 Confusion Matrix .....	60
3.3.2 ROC Chart .....	63
3.3.3 PR-ROC Curve .....	67
3.4 Summary .....	70
<b>CHAPTER FOUR DESIGN AND DEVELOPMENT OF PREDICTIVE MAINTENANCE FRAMEWORK .....</b>	<b>71</b>
4.1 Overview .....	71
4.2 A Proposed Predictive Maintenance Framework .....	72
4.3 The Research Framework.....	74
4.4 Phase I: Theoretical Study.....	76
4.4.1 Structuring Classification Method into Classification Model .....	77
4.5 Phase II: Research Approaches Development.....	81
4.5.1 Choosing Machine-Learning Algorithm .....	81
4.5.2 Data Re-sampling Pseudo-code .....	83
4.5.3 Research Instrument.....	87
4.5.3.1 Setting Rapidminer's Operators.....	87
4.5.3.2 Rapidminer Operator's Set-up .....	109
4.6 Phase III: Experimental Study .....	114
4.6.1 Data Collection .....	114
4.6.2 Data Transformation.....	117



4.6.3 Classification Model.....	122
4.6.3.1 Class Distribution Modulation.....	122
4.6.3.2 Hypothesis Space Augmentation .....	124
4.6.3.3 Voting and Stacking Ensemble .....	126
4.7 Phase IV: Results and Analysis .....	127
4.8 Summary .....	128
<b>CHAPTER FIVE RESULTS AND ANALYSIS .....</b>	<b>129</b>
5.1 Overview .....	129
5.2 Evaluate Skew Datasets .....	129
5.2.1 Experimental Procedure .....	129
5.2.2 Result Analysis .....	131
5.2.2.1 The impact of learner scheme .....	135
5.2.2.2 The impact of bit pattern number.....	138
5.2.2.3 The impact of bit length.....	141
5.2.2.4 The impact of data partition size .....	145
5.3 Optimal Class Distribution, and Re-sampling Technique.....	148
5.3.1 Experimental Procedure .....	148
5.3.2 Data Re-sampling Procedure .....	149
5.3.3 Result Analysis .....	151
5.3.3.1 The impact of learner scheme .....	153
5.3.3.2 The impact of bit pattern number.....	165
5.3.3.3 The impact of bit length.....	167
5.3.3.4 The impact of data re-sampling technique.....	170
5.3.3.5 The impact of data re-sampling ratio .....	174
5.3.3.6 The impact of data partition size .....	180
5.3.3.7 The impact of class ratio in data re-sampling .....	186
5.4 Optimal Ensemble Learner for Hypothesis Augmentation with Bagging and Boosting .....	197
5.4.1 Experimental Procedure .....	197
5.4.2 Result Analysis .....	197

5.5 Optimal Ensemble Learner for Hypothesis Augmentation with Voting and Stacking.....	206
5.5.1 Experimental Procedure .....	206
5.5.2 Result Analysis .....	208
5.6 Summary of Findings.....	216
5.7 The Proposed Predictive Maintenance Framework .....	218
5.8 Validation of Proposed Prediction Model through Comparison study with other researcher's work.....	219
5.8.1 Experimental procedure.....	221
5.8.2 Result Analysis .....	221
5.9 Summary .....	231
<b>CHAPTER SIX CONCLUSIONS AND FUTURE WORK .....</b>	<b>233</b>
6.1 Overview .....	233
6.2 Discussions .....	234
6.3 Implications of Research .....	235
6.4 Research Limitations.....	237
6.5 Research Contribution.....	238
6.5.1 Methodological Contributions .....	239
6.5.2 Theoretical Contribution.....	240
6.5.3 Experimental Contribution .....	240
6.6 Publications of the Research .....	242
6.7 Future Work.....	243
<b>REFERENCES .....</b>	<b>246</b>

## List of Tables

Table 1.1: Research Objectives .....	6
Table 1.2: The Research Approach .....	7
Table 1.3: Theoretical Study .....	8
Table 1.4: Research Framework Development .....	9
Table 1.5: Experimental Study .....	10
Table 1.6: Confirmatory Study .....	11
Table 1.7: Terms and its descriptions .....	12
Table 4.1: KStar learning scheme descriptions and settings .....	88
Table 4.2: IBk learning scheme descriptions and settings .....	89
Table 4.3: LWL learning scheme descriptions and settings .....	90
Table 4.4: J48 learning scheme descriptions and settings .....	91
Table 4.5: Naïve Bayes learning scheme descriptions and settings .....	92
Table 4.6: Boosting learning scheme descriptions and settings .....	93
Table 4.7: Bagging learning scheme descriptions and settings .....	94
Table 4.8: Vote learning scheme descriptions and settings .....	95
Table 4.9: Stacking learning scheme descriptions and settings .....	96
Table 4.10: Principle Component Analysis (PCA) descriptions and settings .....	97
Table 4.11: Correlation Matrix descriptions and settings .....	98
Table 4.12: Genetic algorithm for feature selection descriptions and settings .....	99
Table 4.13: Outliers identifiers descriptions and settings .....	102
Table 4.14: Weight by correlation descriptions and settings .....	103
Table 4.15: Selection by weight descriptions and settings .....	104
Table 4.16: Cross-Validation descriptions and settings .....	105
Table 4.17: Binomial classification performance evaluator descriptions and settings .....	106
Table 4.18: T-Test performance evaluator descriptions and settings .....	108
Table 4.19: Example of the collected data .....	115
Table 4.20: Details of attributes .....	116
Table 4.21 The selected classifiers for Voting and Stacking learner .....	127

Table 5.1: Sample of discretized data through quinary bit number with 8 bit length .....	131
Table 5.2: Imbalance data sets training result .....	132
Table 5.3: Imbalance data sets test result.....	134
Table 5.4: P&R result on D153 with under-sampling class ratios .....	189
Table 5.5: P&R result on D153 with under-sampling Kubat class ratios.....	190
Table 5.6: P&R result on D153 with SMOTE under-sampling class ratios .....	191
Table 5.7: P&R result on D153 with SMOTE under-sampling Kubat class ratios .	192
Table 5.8: P&R result on D153 with SMOTE class ratios .....	193
Table 5.9: P&R result on D153 with over-under-sampling class ratios .....	194
Table 5.10: P&R result on D153 with over-sampling class ratios .....	195
Table 5.11: T-Test between SMOTE-Undersampling with Naïve Bayes Quinary PB8 and Imbalance data sets test result.....	196
Table 5.12: Selected datasets for bagging, boosting and combination of both.....	198
Table 5.13: T-Test result between boosting & bagging and SMOTE-Undersampling with Naïve Bayes Quinary PB8.....	205
Table 5.14: Selected datasets based from D153 test result .....	208
Table 5.15: Selected datasets based from D154 test result .....	209
Table 5.16: AUC result for strong model based from Table 5.14.....	209
Table 5.17: AUC result for weak model based from Table 5.14 .....	210
Table 5.18: AUC result for weakstrong model based from Table 5.14.....	210
Table 5.19: AUC result for strong model based from Table 5.15 .....	211
Table 5.20: AUC result for weak model based from Table 5.15 .....	211
Table 5.21: AUC result for weakstrong model based from Table 5.15.....	212
Table 5.22: Selected diverse datasets for voting and stacking based from AUC-PR .....	213
Table 5.23: Voting and Stacking test result .....	214
Table 5.24: T-Test result between boosting & bagging and voting & stacking test result.....	215
Table 5.25: Summary of research objectives and its respective outcome. ....	216
Table 5.26: Training results of the Comparison#1 prediction model.....	221
Table 5.27: Training results of the Comparison#2 prediction model.....	222

Table 5.28: Training results of the proposed prediction model .....	222
Table 5.29: Test results of the Comparison#1 prediction model .....	225
Table 5.30: Test results of the Comparison#2 prediction model .....	226
Table 5.31: Test results of the proposed prediction model .....	226
Table 5.32: T-Test results between comparison#1 and comparison#2.....	229
Table 5.34: T-Test between test results of comparison#2 and SMOTE	
Undersampling through Naive Bayes, Quinary, & PB8.....	231
Table 6.1: Contribution and its respective publication .....	242
Table 6.2: Other additional contributions .....	243

## List of Figures

Figure 1.1: Thesis Organization .....	17
Figure 1.2: Conceptual Framework .....	19
Figure 2.1: Classification of maintenance (Misra, 2008).....	22
Figure 2.2: Comparison#1 prediction model (Zhou et al., 2005).....	31
Figure 2.3: Comparison#2 Prediction Model (Collier & Held, 2000).....	32
Figure 2.4: Increases in training data affects performance (Weiss G. M., 2003).....	43
Figure 3.1: Training in classification methodology (Lanzi, 2006).....	54
Figure 3.2 Testing in classification methodology (Lanzi, 2006).....	54
Figure 3.3: Tomek Link procedure flow chart (Kubat & Matwin, 1997).....	56
Figure 3.4: CNN procedure pseudo code (Kubat & Matwin, 1997) .....	57
Figure 3.5 Evaluation of classifier model accuracy.....	59
Figure 3.6: Confusion Matrix.....	61
Figure 3.7: ROC trade-off between TPrate and FPrate (Flach, 2003).....	64
Figure 3.8: ROC curve ideal point (Caruana, 2004).....	65
Figure 3.9: PR curve trade-off between Precision and Recall (Caruana, 2004) .....	68
Figure 4.1: Proposed Predictive Maintenance Framework .....	73
Figure 4.2: Research Framework Flow.....	75
Figure 4.3: The Classification Model of Time-series Rare Event.....	78
Figure 4.4: Random under-sampling pseudo-code.....	83
Figure 4.5: SMOTE algorithm pseudo-code .....	84
Figure 4.6: Create Synthetics Instance pseudo-code .....	84
Figure 4.7: Random over-sampling pseudo-code.....	85
Figure 4.8: Kubat under-sampling pseudo-code .....	85
Figure 4.9: Tomek Link algorithm pseudo-code.....	86
Figure 4.10: Condensed Nearest Neighbour pseudo-code.....	86
Figure 4.11: Training operator's set-up .....	109
Figure 4.12: Prediction model testing operator's set-up.....	110
Figure 4.13: Operator's set-up within Cross-Validation operator.....	110
Figure 4.14: Learner operators' set-up in voting operator .....	111
Figure 4.15: Learner operators' set-up in stacking operator .....	111

Figure 4.16: Comparison#1 operators' set-up.....	112
Figure 4.17: Comparison#2 operators' set-up.....	112
Figure 4.18: Operator's set-up within Cross-Validation operator.....	113
Figure 4.19: T-Test operators' set-up .....	113
Figure 4.20: Yield output and attributes trend. ....	119
Figure 4.21: Experimental conditions on skewed datasets evaluation .....	123
Figure 4.22: The experimental conditions for Ensemble Learner.....	125
Figure 5.1: Precision and Recall training result for Imbalance data sets.....	133
Figure 5.2: Precision and Recall test result for imbalance data sets.....	135
Figure 5.3: P&R result on D139 with different type of learner scheme.....	136
Figure 5.4: P&R result on D153 with different type of learner scheme.....	137
Figure 5.5: P&R result on D154 with different type of learner scheme.....	138
Figure 5.6: P&R result on D139 with different type of bit pattern number.....	139
Figure 5.7: P&R result on D153 with different type of bit pattern number.....	140
Figure 5.8: P&R result on D154 with different type of bit pattern number.....	140
Figure 5.9: P&R result on D139 with different bit pattern length in quinary .....	142
Figure 5.10: P&R result on D153 with different bit pattern length in quinary .....	142
Figure 5.11: P&R result on D154 with different bit pattern length in quinary number .....	143
Figure 5.12: P&R result on D139 with different bit pattern length in binary number .....	144
Figure 5.13: P&R result on D153 with different bit pattern length in binary number .....	144
Figure 5.14: P&R result on D154 with different bit pattern length in binary number .....	145
Figure 5.15: P&R result on D139 with different partition size in binary number...	146
Figure 5.16: P&R result on D153 with different partition size in binary number...	147
Figure 5.17: P&R result on D154 with different partition size in binary number...	147
Figure 5.18: Overall re-sample data sets training result .....	152
Figure 5.19: Overall P&R test result of re-sample model.....	152
Figure 5.20: P&R training result with IBK of learner scheme.....	154
Figure 5.21: P&R training result with different J48 of learner scheme.....	154

Figure 5.22: P&R training result with different $K^*$ of learner scheme .....	155
Figure 5.23: P&R training result with different LWL of learner scheme.....	155
Figure 5.24: P&R training result with different Naïve Bayes of learner scheme ...	156
Figure 5.25: P&R result on D139 with IBK learning scheme .....	156
Figure 5.26: P&R result on D139 with J48 learning scheme.....	157
Figure 5.27: P&R result on D139 with $K^*$ learning scheme .....	157
Figure 5.28: P&R result on D139 with LWL learning scheme.....	158
Figure 5.29: P&R result on D139 with Naïve Bayes learning scheme.....	158
Figure 5.30: P&R result on D153 with IBK learning scheme .....	159
Figure 5.31: P&R result on D153 with J48 learning scheme.....	159
Figure 5.32: P&R result on D153 with $K^*$ learning scheme .....	160
Figure 5.33: P&R result on D153 with LWL learning scheme.....	160
Figure 5.34: P&R result on D153 with Naïve Bayes learning scheme.....	161
Figure 5.35: P&R result on D154 with IBK learning scheme .....	161
Figure 5.36: P&R result on D154 with J48 learning scheme.....	162
Figure 5.37: P&R result on D154 with $K^*$ learning scheme .....	163
Figure 5.38: P&R result on D154 with LWL learning scheme.....	163
Figure 5.39: P&R result on D154 with Naïve Bayes learning scheme.....	164
Figure 5.40: P&R result on D139 in relation with binary and quinary number.....	166
Figure 5.41: P&R result on D153 in relation with binary and quinary number.....	166
Figure 5.42: P&R result on D154 in relation with binary and quinary number.....	167
Figure 5.43: P&R result on D139 in relation with bit length.....	168
Figure 5.44: P&R result on D153 in relation with bit length.....	169
Figure 5.45: P&R result on D154 in relation with bit length.....	169
Figure 5.46: P&R result on D139 with different sampling technique .....	171
Figure 5.47: P&R result on D139 with different sampling technique (without US).....	172
Figure 5.48: P&R result on D153 with different sampling technique .....	172
Figure 5.49: P&R result on D153 with different sampling technique (without US).....	173
Figure 5.50: P&R result on D154 with different sampling technique .....	173
Figure 5.51: P&R result on D154 with different sampling technique (without US).....	174
Figure 5.52: P&R result on D139 with SMOTE sampling ratios.....	176
Figure 5.53: P&R result on D139 with SMOTE-KUS sampling ratios.....	176



Figure 5.54: P&R result on D139 with SMOTE-US sampling ratios .....	177
Figure 5.55: P&R result on D153 with SMOTE sampling ratios.....	177
Figure 5.56: P&R result on D153 with SMOTE-KUS sampling ratios.....	178
Figure 5.57: P&R result on D153 with SMOTE-US sampling ratios .....	178
Figure 5.58: P&R result on D154 with SMOTE sampling ratios.....	179
Figure 5.59: P&R result on D154 with SMOTE-KUS sampling ratios.....	179
Figure 5.60: P&R result on D154 with SMOTE-US sampling ratios .....	180
Figure 5.61: P&R result on D139 with SMOTE-US sampling and partition ratios	182
Figure 5.62: P&R result on D153 with SMOTE-US sampling and partition ratios	182
Figure 5.63: P&R result on D154 with SMOTE-US sampling and partition ratios	183
Figure 5.64: P&R result on D139 with SMOTE-KUS sampling and partition ratios .....	183
Figure 5.65: P&R result on D153 with SMOTE-KUS sampling and partition ratios .....	184
Figure 5.66: P&R result on D154 with SMOTE-KUS sampling and partition ratios .....	184
Figure 5.67: P&R result on D139 with SMOTE sampling and partition ratios .....	185
Figure 5.68: P&R result on D153 with SMOTE sampling and partition ratios .....	185
Figure 5.69: P&R result on D154 with SMOTE sampling and partition ratios .....	186
Figure 5.70: Boosting test result with D139 .....	199
Figure 5.71: Bagging test result with D139 .....	200
Figure 5.72: Bagging & Boosting test result with D139 .....	200
Figure 5.73: Boosting & Bagging test result with D139 .....	201
Figure 5.74: Boosting test result with D153 .....	201
Figure 5.75: Bagging test result with D153 .....	202
Figure 5.76: Bagging & Boosting test result with D153 .....	202
Figure 5.77: Boosting & Bagging test result with D153 .....	203
Figure 5.78: Boosting test result with D154 .....	203
Figure 5.79: Bagging test result with D154 .....	204
Figure 5.80: Bagging & Boosting test result with D154 .....	204
Figure 5.81: Boosting & Bagging test result with D154 .....	205
Figure 5.82: Proposed Predictive Maintenance Framework for Manufacturing.....	218

Figure 5.83: Prediction model flow for comparison #1 (Zhou et al. (2005)).....	220
Figure 5.84: Prediction model flow for comparison #2 (Collier & Held (2000)) ...	220
Figure 5.85: Proposed prediction model flow .....	220
Figure 5.86: P&R Curve for training results of Comparison#1 .....	223
Figure 5.87: P&R Curve for training results of Comparison#2 .....	224
Figure 5.88: P&R Curve for training results of proposed prediction model.....	224
Figure 5.89: P&R Curve for test results of Comparison#1 .....	227
Figure 5.90: P&R Curve for test results of Comparison#2.....	228
Figure 5.91: P&R Curve for test results of proposed prediction model .....	228

## **List of Appendices**

Appendix 1 Date Range and Number of Datasets .....	262
Appendix 2 Re-sampled Datasets Induced Classifiers .....	263
Appendix 3 Bagging, Boosting Induced Classifier Models .....	269
Appendix 4 Imbalance Datasets Training Result .....	270
Appendix 5 Imbalance Datasets Test Result .....	272
Appendix 6 Overall Re-sampled Datasets Training Result .....	274
Appendix 7 T-Test between SMOTE Under-sampling Naïve Bayes Quinary PB8 and Imbalance Datasets .....	276
Appendix 8 T-Test Result between Boosting & Bagging and SMOTE Under-sampling of Quinary PB8 .....	277
Appendix 9 T-Test Result between Boosting & Bagging and Voting & Stacking Test Result .....	278
Appendix 10 T-Test between Comparisons#1 and Comparisons#2 Test Result ....	279
Appendix 11 T-Test between Comparisons#1 and SMOTE Under-sampling Naive Bayes Quinary .....	280
Appendix 12 T-Test between Comparisons#2 and SMOTE Under-sampling Naïve Bayes Quinary PB8 .....	281

## **List of Abbreviations**

AAv	- Above Average
BAv	- Below Average
CL	- Control Limit
CNN	- Condensed Nearest Neighbour
FN	- False Negative
FP	- False Positive
FEM	- Fuji Electric Malaysia
G+R %	- Glide + Read/Write Percentage
G-NG	- Glide No Good
IID	- Independent and Identically Distributed
IPA	- Isopropyl-Alcohol
K-NN	- K-Nearest Neighbour
MDL	- Minimum Description Length
MP	- Missing Pulse
NPV	- Negative Predictive Value
OS	- Random Over-sampling,
OSUS	- Random Over-sampling & Under-sampling,
PPV	- Positive Predictive Value,
PR	- Precision and Recall
RIE	- Reactive Ion Etching
ROC	- Receiver Operating Characteristic
R-NG	- Read/Write No Good
SMART	- Self Monitoring and Reporting Technology
SMK	- SMOTE & Kubat under-sampling
SMT	- SMOTE
SMU	- SMOTE & Random Under-sampling
SR	- Sampling Ratio
SMOTE	- Synthetics Minority Oversampling Technique
AUC-PR	- Total Area Under the PR Curve
AUC-ROC	- Total Area Under the ROC Curve

TN	- True Negative
TP	- True Positive
US	- Random Under-sampling
USK	- Kubat Under-sampling
VC	- Vapnik-Chervonenkis
YD	- Yeild Drop

# **CHAPTER ONE**

## **INTRODUCTION**

### **1.1 Research Background**

Technology based manufacturers such as hard disk media and semiconductor have been putting efforts to reduce an operation cost in manufacturing process (Gardner & Bieker, 2000; Halevi, 2006). For instance, hard disk media manufacturers are facing challenges to produce higher densities disk media with a cheaper operation cost (Ricker, 2007). In line with this fact, the trend of implementing predictive maintenance has become one of an important option for manufacturers to increase their business competitiveness (Yang, Djurdjanovic, & Ni, 2007; Zhou, et al., 2005; Zeng, et al., 2006; Lin & Tseng, 2005). This has led to a need of predictive maintenance framework.

Various efforts have being done by many researchers to produce prediction models (Tse P. W., 2002; Zhou, et al., 2005; Sodiya, 2005; Himmel, Kim, Krauss, Kamen, & May, 1995; Choi & Kim, 1999; Kim, et al., 2000; Dupret & Kielbasa, 2004; Rietman, 1997; Hughes, Murray, Kreutz D., & Elkan, 2002; Pinheiro, Weber, & Barroso, 2007). However, most of the mentioned researchers have focused on issues such as vibration, temperature or other machine health nature of data analysis that are logged from equipments to produce prediction model. These models have been developed using two approaches, learning from equipments' mechanism behaviours and process parameters. The former approach is intended to predict equipment failures, and the latter is to determine optimal process settings. In the context of

The contents of  
the thesis is for  
internal user  
only

## REFERENCES

- Abe, H., Ohsaki, M., Yokoi, H., & Yamaguchi, T. (2006). *New Frontiers in Artificial Intelligence*. Springer Berlin / Heidelberg.
- Alpaydin, E. (2010). *Introduction to Machine Learning*. The MIT Press,.
- Aÿfalg, J. (2008). Advance Analysis on Temporal Data. *Dissertation of Mathematik, Informatik and Statistik, Munchen* .
- Bach, F. R., Heckerman, D., & Horvitz, E. (2005). On the path to an ideal ROC curve: Considering cost asymmetry in learning classifiers., (pp. 9-16).
- Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.* , 6 (1), 20-29.
- Bauer, E., & Kohavi, R. (1999). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning* , 36 (1), 105-139.
- Beck, J. R., & Schultz, E. K. (1986.). The use of ROC curves in test performance evaluation. *Arch Pathol* , 110:13–20.
- Begg, C. D., Merdes, T., Byington, C., & Maynard, K. (1999). Dynamics Modeling for Mechanical Fault Diagnostics and Prognostics. *Proceedings of MARCON 99, Maintenance and Reliability Conference* (p. 22). Gatlinburg, Tennessee.
- Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kotter, T., Meinl, T., et al. (2007). KNIME: The Konstanz Information Miner. *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer.
- Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1987). Occams Razor. *Inf. Process. Lett.* , 24, 377-380.
- Bockhorst, J., & Craven, M. (2005). Markov networks for detecting overlapping elements in sequence data. *Neural Information Processing Systems 17*. MIT Press.
- Bousquet, O., Boucheron, S., & Lugosi, G. (2004). Introduction to Statistical Learning Theory. (pp. 169--207). Springer.
- Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* , 30, 1145 - 1159.



- Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science* .
- Breiman, L., & Spector, P. (1992). Submodel selection and evaluation in regression: The X-random case. *International Statistical Review* , 291-319.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall, New York, NY.
- Bunescu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K., et al. (2004). Comparative Experiments on Learning Information Extractors for Proteins and their Interactions.
- Campos, M., Palma, J., & Marín, R. (2007). Temporal Data Mining with Temporal Constraints. *4594*, pp. 67-76.
- Cardie, C., & Howe, N. (1997). Improving Minority Class Prediction Using Case-Specific Feature Weights. (pp. 57-65). Morgan Kaufmann.
- Caruana, R. (2004, September 15 ). *Performance Measures*. Retrieved October 20, 200, from cs.cornell.edu:  
[http://www.cs.cornell.edu/Courses/cs578/2003fa/performance\\_measures.pdf](http://www.cs.cornell.edu/Courses/cs578/2003fa/performance_measures.pdf)
- Chan, K.-P., & Fu, A. W. (1999). Efficient Time Series Matching by Wavelets., (pp. 126-133).
- Chan, P. K., & Stolfo, S. J. (1998). Toward Scalable Learning with Non-uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. *Knowledge Discovery and Data Mining* (pp. 164-168). AAAI Press.
- Chapple, M. (2010, June 11). *Data Mining: An Introduction*. Retrieved September 20, 2010, from databases.about.com:  
<http://databases.about.com/od/datamining/a/datamining.htm>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* , 16, 321-357.
- Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: special issue on learning from imbalanced datasets. *SIGKDD Explor. Newsl.* , 6, 1-6.
- Chew, C. (2006). *Media Defect Library*. Fuji Electric Malaysia.
- Choi, M.-K., & Kim, H.-M. (1999). A Study on Process Control to Improve Yield in Semiconductor Manufacturing., (p. 1215-1219).

- Cieslak, D. A., Chawla, N. V., & Striegel, A. (2006). Combating imbalance in network intrusion datasets. *In Proceedings of the IEEE International Conference on Granular Computing*, (pp. 732–737).
- Cohen, G., Hilario, M., Sax, H., Hugonnet, S., & Geissbuhler, A. (2006). Learning from imbalanced data in surveillance of nosocomial infection. *Artificial Intelligence in Medicine*, 37(1):7–18.
- Cohen, P. R. (1995). Empirical methods for artificial intelligence. *IEEE Expert: Intelligent Systems and Their Applications*, Vol 11 issue 6 pp 88.
- Collier, K., & Held, G. (2000). Data Mining Quality in Manufacturing Data. *SAS & KPMG Consulting*, (Best Practices Approach To The Manufacturing Industry), 24.
- Compumine. (2008). *Evaluating a classification model – What does precision and recall tell me?* Retrieved April 14, 2010, from Compumine:  
<http://www.compumine.com/web/public/newsletter/20071/precision-recall>
- Cotofrei, P., & Stoffel, K. (2005). First-Order Logic Based Formalism for Temporal Data Mining. In T. Y. Lin, S. Ohsuga, C.-J. Liao, X. Hu, & S. Tsumoto, *Foundations of Data Mining and knowledge Discovery* (Vol. 6/2005, pp. 185-210). Springer Berlin / Heidelberg.
- Das, G., Lin, K.-i., Mannila, H., Renganathan, G., & Smyth, P. (1998). Rule Discovery From Time Series. *KDD* (pp. 16-22). AAAI Press.
- Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning* (pp. 233-240). Pittsburgh, Pennsylvania: ACM.
- Davis, J., Burnside, E., Dutra, I., Page, D., Ramakrishnan, R., Costa, V. S., et al. (2005). View learning for statistical relational learning: with an application to mammography. *Proceedings of the 19th international joint conference on Artificial intelligence* (pp. 677--683). Edinburgh, Scotland: Morgan Kaufmann Publishers Inc.
- Devroye, L., Györfi, L., & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. New-York: Springer–Verlag.
- Dietterich, T. G. (1997). Machine Learning Research: Four Current Directions. *AI Magazine*, 18, 97-136.
- Divyakant, Y.-L. W., Agrawal, D., & Abadi, A. E. (2000). A Comparison of DFT and DWT Based Similarity Search in Time-Series Databases., (pp. 488-495).
- Donkers, H. (2006, July 3). *Data mining and the knowledge discovery process*. Retrieved October 20, 2010, from cs.unimaas.nl:

- Drummond, C., & Holte, R. C. (2000). Explicitly Representing Expected Cost: An Alternative to ROC Representation. *In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 198--207). ACM Press.
- Drummond, C., & Holte, R. (2003). C4.5, class imbalance, and cost sensitivity: why under-sampling beats oversampling. *n Proc. Workshop on Learning from Imbalanced Datasets II*, (pp. 1-8).
- Dupret, Y., & Kielbasa, R. (2004). Modeling Semiconductor Manufacturing Yield by Test Data and Partial Least Squares. *16th International Conference on Microelectronics* (pp. 134-137). Tunisia: IEEE.
- Estad Software. (2010). *What are the most used data mining techniques?* Retrieved December 1, 2011, from <http://www.apa.org/ethics/code.html>
- Estabrooks, A., & Japkowicz, N. (2001). A Mixture-of-Experts Framework for Learning from Imbalanced Datasets. *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis. 2189*, pp. 34-43. London, UK: Springer Berlin / Heidelberg.
- Evgeniou, T., Pontil, M., & Poggio, T. (2000). Statistical Learning Theory: A Primer. *Int. J. Comput. Vision* , 38 (1), 9-13.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.* , 27 (8), 861-874.
- Fayyad, U. M. (1997). Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases. *Proceedings of the Ninth International Conference on Scientific and Statistical Database Management* (pp. 2-11). Washington, DC, USA: ACM.
- Flach, P. A. (2003). The geometry of ROC space: understanding machine learning metrics through ROC isometrics. *in Proceedings of the Twentieth International Conference on Machine Learning* (pp. 194-201). AAAI Press.
- Freund, Y., Györfi, L., Turán, G., & Zeugmann, T. (2008). Algorithmic Learning Theory. *In Proceedings of 19th International Conference, ALT. 5254*. Budapest, Hungary: Springer.
- Fukunaga, K. (1990). *Introduction to statistical pattern recognition (2nd ed.)*. San Diego, CA, USA: Academic Press Professional, Inc.
- Galushka, M., Patterson, D., & Rooney, N. (2006). Temporal Data Mining for Smart

- Homes. *Designing Smart Homes*, vol.4008/2006 of *Lecture Notes in Computer Science*. 4008/2006, pp. 85-108. Springer Berlin / Heidelberg.
- Gao, S., Lee, C. H., & Lim, J.-H. (2006). An ensemble classifier learning approach to ROC optimization. *18th International Conference on Pattern Recognition* (pp. 679-682). Hong Kong, China: IEEE Computer Society.
- García, S., & Herrera, F. (2009). Evolutionary Undersampling for Classification with Imbalanced Datasets: Proposals and Taxonomy. *Evolutionary Computation*, 17, 275-306.
- Gardner, M., & Bieker, J. (2000). Data mining solves tough semiconductor manufacturing problems. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 376--383). Boston, Massachusetts, United States: ACM.
- Gilad B. (2008, March 02). "*Learning Has Just Started*" - an interview with Prof. Vladimir Vapnik. Retrieved Jan 20, 2010, from learningtheory: [http://www.learningtheory.org/index.php?option=com\\_content&view=article&id=9%3Aqllearning-has-just-startedq-an-interview-with-prof-vladimir-vapnik&catid=12%3Ainterviews&Itemid=8](http://www.learningtheory.org/index.php?option=com_content&view=article&id=9%3Aqllearning-has-just-startedq-an-interview-with-prof-vladimir-vapnik&catid=12%3Ainterviews&Itemid=8)
- Goadrich, M., Oliphant, L., & Shavlik, J. (2004). Learning Ensembles of First-Order Clauses for Recall-Precision Curves: A Case Study in Biomedical Information Extraction. *Proceedings of the 14th International Conference on Inductive Logic Programming*, (pp. 98-115). Porto, Portugal.
- Goh, L. (2003). *QIS System*. techreport.
- Grover, L. K., & Mehra, R. (2008). The Lure of Statistics in Data Mining. *Journal of Statistics Education*, 16.
- Grzymala B., J., Stefanowski, J., & Wilk, S. (2005). A Comparison of Two Approaches to Data Mining from Imbalanced Data. *Journal of Intelligent Manufacturing*, 16 (6), 565-573.
- Gu, Q., Cai, Z., & Zhu, L. (2009). Classification of Imbalanced Datasets by Using the Hybrid Re-sampling Algorithm Based on Isomap. *Proceedings of the 4th International Symposium on Advances in Computation and Intelligence* (pp. 287-296). Huangshi, China: Springer-Verlag.
- Guh, R.-S., Zorriassatine, F., Tannock, J. D. T., & O'Brien, C. (1999). On-line control chart pattern detection and discrimination--a neural network approach. *Artificial Intelligence in Engineering*, 13(4), 413-425. doi: DOI: 10.1016/S0954-1810(99)00022-9.

- Guo, H., & Viktor, H. L. (2004). Learning from imbalanced datasets with boosting and data generation: the DataBoost-IM approach. *SIGKDD Explorations* , 6, 30-39.
- Halevi, G. (2006). *Industrial Competitiveness Cost Reduction*. Springer Netherlands.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations, Volume 11, Issue 1* , Volume 11, Issue 1.
- Hand, D. (1997). *Construction and Assessment of Classification Rules*. Chichester: John Wiley & Sons.
- Hansen, L. K., & Salamon, P. (1990). Neural Network Ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* , 12 (10), 993-1001.
- Hart, P. E. (1968). The Condensed Nearest Neighbor Rule. *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, Vol. 14, pp. 515-516.
- Himmel, C. D., Kim, T. S., Krauss, A., Kamen, E. W., & May, G. S. (1995). real-time predictive control of semiconductor manufacturing process using neural network. in *Proceedings of the 1995 American Control Conference*, 2, pp. 1240-1244.
- Hofemann, P. (2009, January 23). *A collaborative course for HDD*. Retrieved February 11, 2009, from solid-state.com: <http://www.solid-state.com/display/article/349314/5/none/none/Dept/Acollaborative-course-for-HDD-manufacturing>
- Hofemann, P. (2008, September 23). *Hard Disk Drive Industry Driving Areal Density and Lithography*. Retrieved February 11, 2009, from molecularimprints.com: <http://www.molecularimprints.com/NewsEvents/tech/articles/Diskcon/Hofemann2.pdf>
- Horner, R., El-Haram, M., & Munns, A. (1997). Building maintenance strategy: a new management approach. *Journal of Quality in Maintenance Engineering*, 3, 273 - 280.
- Huang, F., Xie, G., & Xiao, R. (2009). Research on Ensemble Learning. *Artificial Intelligence and Computational Intelligence, International Conference on* , 3, 249-252.
- Huang, J., Smola, A. J., Gretton, A., Borgwardt, K. M., & Scholkopf, B. (2007).

- Correcting Sample Selection Bias by Unlabeled Data. *In Advances in neural information processing systems 19*, (pp. 601-608).
- Hughes, G. F., Murray, J. F., Kreutz-Delgado, K., & Elkan, C. (2002). Improved disk-drive failure warnings. *IEEE Transactions on Reliability* , 51, 350 - 357.
- Information entropy*. (2010, February 12). Retrieved September 23, 2010, from worldlingo.com:  
[http://www.worldlingo.com/ma/enwiki/en/Information\\_entropy](http://www.worldlingo.com/ma/enwiki/en/Information_entropy)
- Inman, R. A. (2010, February 12). *Maintenance*. Retrieved November 25, 2010, from referenceforbusiness.com:  
<http://www.referenceforbusiness.com/management/LogMar/Maintenance.html>
- Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intell. Data Anal.* , 6, 429-449.
- Jensen, D. D., & Cohen, P. R. (2000). Multiple Comparisons in Induction Algorithms. *Mach. Learn.* , 38 (3), 309-338.
- John, L. (2002). *Media Defect Summary*. Fuji Electric Malaysia.
- Joshi, M. V., P. Kumar, V., & Agarwal, R. C. (2001). Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements. *Proceedings of the 2001 IEEE International Conference on Data Mining* (pp. 257-264). Washington, DC, USA: IEEE Computer Society.
- Kam, P.S., & Fu, A. W. (2000). Discovering Temporal Patterns for Interval-Based Events. *Proceedings of the Second International Conference on Data Warehousing and Knowledge Discovery* (pp. 317--326). London, UK: Springer-Verlag.
- Kang, S., Kim, J., Chae, J., Choi, W., & Lee, S. (2007). Similarity search using the polar wavelet in time series databases. *Proceedings of the intelligent computing 3<sup>rd</sup> international conference on Advanced intelligent computing theories and applications* (pp. 1347-1354). Qingdao, China: Springer-Verlag.
- Karagiannopoulos, M. G., Anyfantis, D. S., Kotsiantis, S. B., & Pintelas, P. E. (2007). Local cost sensitive learning for handling imbalanced datasets. *Mediterranean Conference on Control & Automation* , 1-6.
- Kempe, S., Hipp, J., & Kruse, R. (2008). Data Analysis, Machine Learning and Applications. In F. A. Temporal, *Data Analysis, Machine Learning and Applications* (pp. 253-260). Springer Berlin Heidelberg.

- Keogh, E., Chakrabarti, K., Pazzani, M., & Mehrotra, S. (2001). Locally adaptive dimensionality reduction for indexing large time series databases. *SIGMOD Rec.*, 30 (2), 151-162.
- Kim, O., & Kasmer, L. (2006). What Is Prediction and What Can Prediction Do to Promote Reasoning? *The annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*.
- Kim, T. S., Ahn, S. H., Jang, Y. G., Lee, J. I., Lee, K. J., Kim, B. Y., et al. (2000). Yield prediction models for optimization of high-speed micro-processor manufacturing processes. *Electronics Manufacturing Technology Symposium, 2000. Twenty-Sixth IEEE/CPMT International* (pp. 368 - 373). Santa Clara, CA, USA : IEEE.
- Kittler, R., & Wang, W. (2000). Data mining for yield improvements. *In Proc. International Conference Modeling Analysis of Semiconductor Manufacturing (MASM)*, (pp. 270-277).
- Knierim, T. (2010, June 10). *Uncertainty Principle*. Retrieved November 25, 2010, from thebigview.com: <http://www.thebigview.com/spacetime/uncertainty.html>
- Koc, M., & Lee, J. (2001). A system framework for next-generation e-maintenance system. *Proceeding of Second International Symposium on Environmentally Conscious Design and Inverse Manufacturing*. Tokyo.
- Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2* (pp. 1137-1145). Montreal, Quebec, Canada: Morgan Kaufmann.
- Kok, S., & Domingos, P. (2005). Learning the structure of Markov logic networks. *Proceedings of the 22nd international conference on Machine learning* (pp. 441-448). Bonn, Germany: ACM.
- Korn, F., Jagadish, H. V., & Faloutsos, C. (1997). Efficiently supporting ad hoc queries in large datasets of time sequences. *SIGMOD Rec*, 26 (2), 289-300.
- Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering VOL.30*.
- Kubat, M., & Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *In Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 179-186). Morgan Kaufmann.
- Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine Learning for the Detection of Oil Spills in Satellite Radar Images. *Mach. Learn.*, 30 (2-3), 195-215.

- Kusiak, A., & Burns, A. (2005). Mining Temporal Data: A Coal-Fired Boiler Case Study. In *Proceedings of the 9th International Conference, KES* (pp.953-958). Melbourne, Australia: Springer.
- Lamberson, P. (2008, April 2). *Inductive Reasoning and Bounded Rationality*. Retrieved November 26, 2010, from mit.edu:  
[http://www.mit.edu/~pjl/page2/files/bounded\\_rationality.pdf](http://www.mit.edu/~pjl/page2/files/bounded_rationality.pdf)
- Landgrebe, T. C., Paclik, P., & Duin, R. P. (2006). Precision-recall operating characteristic (P-ROC) curves in imprecise environments. *Proceedings of the 18th International Conference on Pattern Recognition - Volume 04* (pp. 123-127). Washington, DC, USA: IEEE Computer Society.
- Langford, J. (2005). Tutorial on Practical Prediction Theory for Classification. *J. Mach.Learn. Res.* , 6, 273-306.
- Lanzi, P. L. (2006, November 25). *Machine Learning and Data Mining: 10 Introduction to Classification*. Retrieved November 29, 2010, from slideshare.net:  
<http://www.slideshare.net/pierluca.lanzi/machine-learning-and-data-mining-10-introduction-to-classification>
- Laxman, S., & Sastry, P. S. (2006). A survey of temporal data mining. *SADHANA, Academy Proceedings in Engineering Sciences*.
- Lee, J., & Wang, H. (2008). New Technologies for Maintenance. In *Complex System Maintenance Handbook* (pp. 49-78). Springer London.
- Lewis, D. D., & Gale, W. A. (1994). A Sequential Algorithm for Training Text Classifiers. *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 3-12). Dublin, Ireland: Springer-Verlag.
- Li, C. J., & Yoo, J. (1998). Prognosis of Gear Tooth Crack Growth. *Proceedings of the 52nd Meeting of the Society of Mechanical Failures Prevention Technology* (pp. 419-428). Virginia Beach, VA.
- Lin, C.C., & Tseng, H.-Y. (2005). A neural network application for reliability modelling and condition-based predictive maintenance. *The International Journal of Advanced Manufacturing Technology* , 25, 174-179.
- Ling, C. X., & Li, C. (1998). Data Mining for Direct Marketing: Problems and Solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)* (pp. 73-79). AAAI Press.



- Liu, A., Martin, C., Cour, B., & Ghosh, J. (2010). Effects of Oversampling Versus Cost-Sensitive Learning for Bayesian and SVM Classifiers. In R. a. Stahlbock, *Data Mining in Annals of Information Systems, Volume 8, Part 3* (Vol. 8, pp. 159-192). Springer US.
- Liu, B. (2007). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer.
- Liu, Z., Yu, J. X., Lin, X., Lu, H., & Wang, W. (2005). Locating Motifs in Time-Series Data. In *Advances in Knowledge Discovery and Data Mining* (Vol. 3518/2005, pp. 343-353). Springer Berlin / Heidelberg.
- Lomax, R. G. (2001). *An introduction to statistical concepts*. New Jersey: Lawrence Erlbaum.
- Maloof, M. A. (2003). Learning when datasets are imbalanced and when costs are unequal and unknown. *ICML-Workshop on Learning from Imbalanced Datasets II*.
- Marquez, A. C. (2007). The Maintenance Management Framework. In *Models and Methods for Complex Systems Maintenance* (pp. 69-76). Springer London.
- McCarthy, K., Zabar, B., & Weiss, G. (2005). Does cost-sensitive learning beat sampling for classifying rare classes? *Proceedings of the 1st international workshop on Utility-based data mining* (pp. 69-77). Chicago, Illinois: ACM.
- Mende, T., Koschke, R., & Leszak, M. (2009). Evaluating Defect Prediction Models for a Large Evolving Software System. *Software Maintenance and Reengineering, European Conference on*, 0, 247-250.
- Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8, 283 - 298.
- Mieno, F. S., Shibuya, T., & Odagiri, Y. (1999). Yield improvement using data mining system. *IEEE International Symposium on Semiconductor Manufacturing Conference Proceedings* (pp. 391-394). Santa Clara, CA , USA : IEEE Xplore.
- Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). YALE: rapid prototyping for complex data mining tasks. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 935-940). Philadelphia, PA, USA: ACM.
- Misra, K. B. (2008). *Handbook of Performability Engineering*. Springer London.
- Opitz, D., & Maclin, R. (1999). Popular ensemble methods: an empirical study. *Journal of Artificial Intelligence Research* , 11, 169-198.

- Partridge, D., & Yates, W. B. (1996). Engineering Multiversion Neural-Net Systems. *Neural Computation* , 8, 869-893.
- Payne, T. R., & Edwards, P. (1998). Implicit Feature Selection with the Value Difference Metric. *13th European Conference on Artificial Intelligence* (pp. 450-454). Brighton, UK: John Wiley & Sons.
- Peebles, T. D., Essawy, M. A., & Fein-Sabatto, S. (1999). An intelligent methodology for remaining useful life estimation of mechanical components. *Proceedings of MARCON 99, Maintenance and Reliability Conference* (pp. 27.01-27.09).
- Peter. (2010, September 21 ). *The law of large numbers or the flaw of averages*. Retrieved September 25, 2010, from probabilitytheory.info:  
<http://www.probabilitytheory.info/content/item/6-the-law-of-large-numbers/-the-law-of-averages>
- Phua, C., Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection: classification of skewed data. *SIGKDD Explor. Newsl.* , 6, 50-59.
- Piatetsky-Shapiro, G. (2009, November 6). *Difference between Data Mining and Statistics*. Retrieved October 20, 2010, from kdnuggets.com:  
<http://www.kdnuggets.com/faq/difference-data-mining-statistics.html>
- Pinheiro, E., Weber, W.-D., & Barroso, L. A. (2007). Failure trends in a large disk drive population. *Proceedings of the 5th USENIX conference on File and Storage Technologies* (pp. 2-2). San Jose, CA: USENIX Association.
- Powers, D. M. (2008). Evaluation Evaluation. *18th European Conference on Artificial Intelligence* (pp. 843-844). Patras, Greece: IOS Press.
- Provost, F., & Fawcett, T. (1997). Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions. *In Proceedings of the Third International Conference on Knowledge Discovery and Data Mining* (pp. 43-48). AAAI Press.
- Provost, F., & Fawcett, T. (2001). Robust Classification for Imprecise Environments. *Mach. Learn.* , 42 (3), 203-231.
- Provost, F., Fawcett, T., & Kohavi, R. (1998). The Case Against Accuracy Estimation for Comparing Induction Algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 445-453). San Francisco, CA, USA: Morgan Kaufmann.
- Raskutti, B., & Kowalczyk, A. (2004). Extreme re-balancing for SVMs: a case study. *SIGKDD Explor. Newsl.* , 6 (1), 60-69.

- Richards, G., & Wang, W. (2006). Empirical Investigations on Characteristics of Ensemble and Diversity. *IEEE Proceedings of IJCNN06*.
- Ricker, T. (2007, April 25). *SSD prices in freefall - won't overtake hard disks anytime soon*. Retrieved February 11, 2009, from engadget.com: <http://www.engadget.com/2007/04/25/ssd-prices-in-freefall-wont-overtake-hard-disks-anytime-soon/>
- Rietman, E. A. (1997). Multi Step Process Yield Control with Large System Models. *Proceedings of the 1997 American Control Conference*. 3, pp. 1573-1574. Albuquerque, NM , USA : IEEE.
- Rijsbergen, C. (1979). Information Retrieval. *Journal of the American Society for Information Science* . London: Butterworths: Wiley Subscription Services, Inc., A Wiley Company.
- Russell, S. J., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Pearson Education,.
- Sanchez, P. (2004, 04 29). *Euclidean distance*. Retrieved Mar 20, 2010, from planetmath.org: <http://planetmath.org/encyclopedia/EuclideanDistance.html>
- SAS, I. (2001). *Getting Started With SAS Enterprise Miner*. Cary, NC: SAS Institute Inc.
- Satonori, I. (2002a). *Equipment Maintenance*. Fuji Electric Malaysia.
- Satonori, I. (2002b). *Mechanical Texture and Cleaning*. Fuji Electric Malaysia.
- Satonori, I. (2002c). *Production Process*. Fuji Electric Malaysia.
- Satonori, I. (2003). *Quality Control*. Fuji Electric Malaysia.
- Satonori, I. (2002d). *Tester and Failure Analysis*. Fuji Electric Malaysia.
- Schipper, J. D. (2008, Dec). A knowledge-based toxicology consultant for diagnosing multiple disorders. *Doctorate dissertation* . University of Florida.
- Schölkopf, U. V., & Bernhard. (2009). STATISTICAL LEARNING THEORY: MODELS, CONCEPTS, AND RESULTS.
- Seabhcan. (2005). *Manhattan distance*. Retrieved Feb 05, 2010, from computervision.wikia.com:[http://computervision.wikia.com/wiki/Manhattan\\_distance](http://computervision.wikia.com/wiki/Manhattan_distance)

- Seiffert, C., Khoshgoftaar, T. M., Hulse, J. V., & Napolitano, A. (2008). Improving Learner Performance with Data Sampling and Boosting. *ICTAI (1)* (pp. 452-459). IEEE Computer Society.
- Senin, P. (2008). *Dynamic Time Warping Algorithm Review*. Retrieved from Science. Information and Computer Science Departament University of Hawaii, Honolulu: <http://129.173.35.31/~pf/Linguistique/Treillis/ReviewDTW.pdf>
- Sewell, M. (2008). *Ensemble Learning*. Retrieved Jan 05, 2010, from machine-learning.martinsewell.com: <http://machine-learning.martinsewell.com/ensembles/ensemble-learn>
- Shao, J. (1993). Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association* , 88, 486-494.
- Sheu, D. D., & Kuo, J. Y. (2006). A model for preventive maintenance operations and forecasting. *Journal of Intelligent Manufacturing*, 17, 441-451.
- Shmulevich, I. (2002). *Computational Learning Theory*. Retrieved Feb 09, 2010, from personal.systemsbiology.net: <http://personal.systemsbiology.net/ilya/LEARN.htm>
- Silverman, L. L. (1999). *Critical shift: the future of quality in organizational performance*. *ASQ Quality* (p. 294). Milwaukee: ASQ Quality. Retrieved July 17, 2011, from <http://www.mendeley.com/import/>.
- Singla, P., & Domingos, P. (2005). Discriminative training of markov logic networks. *20th National Conference on Artificial Intelligence* (pp. 868-873). AAAI Press.
- Sodiya, A. S. (2005). Data Mining-based Intelligent Equipment Maintenance. *Journal of Applied Computer Science*, 13, 29-38.
- Stefanowski, J., & Wilk, S. (2006). Rough Sets for Handling Imbalanced Data: Combining Filtering and Rule-based Classifiers . *Fundamenta Informaticae*, 379-391.
- Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, 595–645.
- Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science* , 240, 1285-1293.
- Tadeusz Pietraszek, A. T. (2005). Data Mining and Machine Learning-Towards Reducing False Positives in Intrusion Detection. *Information Security Technical Report Journal*, 10, 169-183.

- Tatavarty, G. (2006). Finding Temporal Association Rules between frequent patterns in Multivariate time series. *Master of Computer Science Thesis*. University of Cincinnati.
- Tavenard, R., Salah, A., & Pauwels, E. (2007). Constructing Ambient Intelligence – AmI. In M. Mühlhäuser, A. Ferscha, & E. Aitenbichler, *AmI 2007 Workshops Darmstadt, Series: Communications in Computer and Information Science, Vol. 11* (Vol. 11, pp. 53-62). Springer Berlin Heidelberg.
- Teknomo, K. (2006, Feb 05). *Chebyshev Distance*. Retrieved Feb 10, 2010, from people.revoledu.com:  
<http://people.revoledu.com/kardi/tutorial/Similarity/ChebyshevDistance.html>
- Tomek, I. (1976). Two modifications of cnn. *IEEE Transactions on Systems, Man, 7*(2), 679–772.
- Tse, P. W. (2002). Maintenance practices in Hong Kong and the use of the intelligent scheduler. *Journal of Quality in Maintenance Engineering*, 8, 369 - 380.
- Tse, P., & Li, L. (2001). Intelligent Predictive Decision Support System for Condition-Based Maintenance. *The International Journal of Advanced Manufacturing Technology*, Volume 17, 383-391.
- Tzanis, G., Kavakiotis, I., & Vlahavas, I. (2009). Innovations in Database Technologies and Applications: Current and Future Trends. In *Handbook of Research on* (pp. 1-1124 pp). IGI Global.
- Vapnik, V. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10, 988-999.
- Waaki, H. (2002). *Error Defect Library*. Fuji Electric Malaysia.
- Wang, D. D., Yang, D., Xu, J., & Xu, K. (1996). Computational intelligence based machine fault diagnosis. *Proceedings of the IEEE International Conference on Industrial Technology, 1996. (ICIT '96)* (pp. 465-469).
- Wang, W. (2008). Some fundamental issues in ensemble methods. *Proceedings of the International Joint Conference on Neural Networks, IJCNN* (pp. 2243-2250). Hong Kong, China: IEEE.
- Wang, W., Jones, P., & Partridge, D. (2000). Diversity between Neural Networks and Decision Trees for Building Multiple Classifier Systems. *Multiple Classifier Systems in Lecture Notes in Computer Science series, 1857*, pp. 240-249.

- Wang, W., Partridge, D., & Etherington, J. (2001 ). Hybrid ensembles and coincident-failure diversity. *International Joint Conference on Neural Networks* (pp. 2376 – 2381 vol.4). Washington, DC, USA : IEEE.
- Watanabe, M. (2002). *GHT Defect Mode*. Fuji Electric Malaysia.
- Webb, B. (1995). *Inductive Learning*. Retrieved Feb 12, 2010, from sifter.org:  
<http://sifter.org/~brandyn/InductiveLearning.html>
- Weiss, G. M. (2003, May). The Effect Of Small Disjuncts And Class Distribution On Decision Tree Learning. *Doctorate Dissertation* . New Brunswick, New Jersey: Rutgers University.
- Weiss, G. M., & Hirsh, H. (1998). Learning to Predict Rare Events in Categorical Time-Series Data. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (pp. 359-363). Menlo Park,CA: AAAI Press.
- Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: the effect of class distribution on tree induction. *J. Artif. Int. Res.* , 19 (1), 315-354.
- Weiss, G. M., McCarthy, K., & Zabar, B. (2007). Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs? In *Proceedings of the 2007 International Conference on Data Mining* (pp. 35-41). CSREA Press.
- Weiss, S. M., & Kulikowski, C. A. (1991). *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. Morgan Kaufmann Publishers Inc.
- Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann.
- Wolpert, D. H. (1992). Stacked Generalization. *Neural Networks*, 5, 241-259.
- Wu, G., & Chang, E. Y. (2003). Class-boundary alignment for imbalanced dataset learning. In *ICML 2003 Workshop on Learning from Imbalanced Datasets*, (pp. 49-56).
- Yam, R. C., Tse, P. W., Li, L., & Tu, P. (2001). Intelligent Predictive Decision Support System for Condition-Based Maintenance. *The International Journal of Advanced Manufacturing Technology*, 17 (5), 383-391.

- Yan, R., Liu, Y., Jin, R., & Hauptmann, A. (2003). On Predicting Rare Classes With Svm Ensembles In Scene Classification. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)* (pp. 21-24). IEEE .
- Yang, Z., Djurdjanovic, D., & Ni, J. (2007). Maintenance scheduling in manufacturing systems based on predicted machine degradation. *Journal of Intelligent Manufacturing*, 19, 87-98.
- Zeng, Y., Jiang, W., Zhu, C., Liu, J., Teng, W., & Zhang, Y. (2006). Prediction of Equipment Maintenance Using Optimized Support Vector Machine. In D.-S. a. Huang, *Computational Intelligence, Lecture Notes in Computer Science series* (Vol. Volume 4114, pp. 570-579). Springer Berlin / Heidelberg.
- Zhang, G., Lee, S., Propes, N., Zhao, Y., Vachtsevanos, G., Thakker, A., et al. (2002). A Novel Architecture for an Integrated Fault Diagnostic/Prognostic System. *AAAI symposium*.
- Zhou, J., Li, X., Andernrooer, A. J., Zeng, H., Goh, K. M., Wong, Y. S., et al. (2005). Intelligent Prediction Monitoring System for Predictive Maintenance in manufacturing. *31st Annual Conference of IEEE in Industrial Electronics Society* (p. 6 pp). IEEE.
- Zhou, Z. H. (2003). Book review: Three perspectives of data mining. *Artif. Intell.*, 143 (1), 139-146.
- Zhou, Z.-H., Wu, J., Tang, W., Zhou, Z.-h., Wu, J., & Tang, W. (2002). Ensembling Neural Networks: Many Could Be Better Than All. *Artif. Intell.* , 239-263.